Article

Not peer-reviewed version

Plan for Constructing DataDiscoveryLab

Elbek Keskinoglu *

Posted Date: 27 April 2023

doi: 10.20944/preprints202304.1074.v1

Keywords: data analysis; computer vision algorithms; visual data; natural language processing; scientific research

Article

Plan for Constructing DataDiscoveryLab

Elbek Javokhir Keskinoglu

Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong SAR, China; elbekjk@gmail.com

Abstract: DataDiscoveryLab is a software tool that enables users to recommend possible pathways to their research with references by extracting valuable insights from academic articles by parsing them into text and figures and processing the image data using computer vision algorithms. The software creates two databases for text-based purposes, one for titles, figure captions, and references, and another for abstracts, introductions, methods, and results using NLP techniques. The software then compares these databases to users' research questions, finds similarities, and presents the findings. Additionally, the software takes data from researchers' scientific software and devices to compare with the current figure-based databases, creating a loop until the best answer and pathways to research and articles to recommend can be found. This tool provides valuable insights and context for researchers, helping them make informed decisions about their research.

Keywords: data analysis; computer vision algorithms; visual data; natural language processing; scientific research

1. Introduction

In the field of research and scientific inquiry, there is a vast amount of information available in the form of academic articles, reports, and studies. This abundance of information can often make it challenging for researchers to find relevant data and gain insights from the existing literature. As a result, there is a growing need for tools that can help researchers effectively parse and analyze the large volumes of data available to them.

DataDiscoveryLab is a software system that addresses this need by providing a comprehensive approach to analyzing academic articles. By utilizing computer vision algorithms and NLP techniques, the system is capable of parsing articles into text and figures, creating two separate databases that can be used to find similarities between users' research questions and existing literature. The system then utilizes this data to provide researchers with the most relevant pathways to research and articles to read, ensuring that they can make informed decisions about their work.

Overall, DataDiscoveryLab is a powerful tool that can help researchers save time and resources while gaining valuable insights from existing literature. With the ability to parse large volumes of data and find relevant information quickly, this system can significantly enhance the research process, ultimately leading to more impactful research outcomes.

2. Article Database and Sub Databases Based on Article Database

The creation of the article database is a crucial part of the DataDiscoveryLab software system. This database is created by collecting articles from different scientific journals and parsing them into text and figures. The text and figures are then analyzed using NLP and computer vision algorithms to create three separate databases.

The first database includes titles, captions, and references from the articles. This database allows researchers to quickly identify articles that may be relevant to their research question. The second database includes abstracts, introductions, methods, and results from the articles. This database provides researchers with more detailed information about the articles and can be used to gain a deeper understanding of the research presented in the article.

The third database is an image-based database that contains the figures from the articles. This database is created by using computer vision algorithms to detect and extract individual figures from the articles. This database can be used to gain insights into the research presented in the articles and can be particularly useful for researchers working in fields that rely heavily on visual data.

In short, the creation of these three databases provides researchers with a comprehensive approach to analyzing academic articles. By utilizing both text-based and image-based data, researchers can gain a more complete understanding of the research presented in the articles. Additionally, the use of NLP and computer vision algorithms allows for quick and efficient analysis of large volumes of data, making it easier for researchers to find relevant information and make informed decisions about their research.

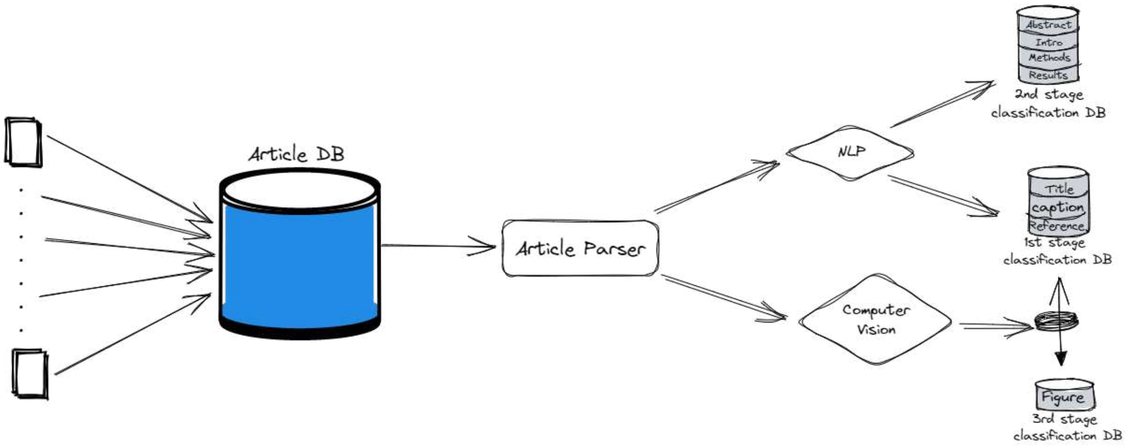

Figure 1. Article Database and sub-databases nucleation.

3. Prompt-Based Search

Once researchers have submitted their research questions to DataDiscoveryLab, the system utilizes LLMs (Language Model Machines) to analyze and understand the questions. This analysis allows the system to find connections between the research questions and the article database that contains titles, captions, and references. These connections are used to identify relevant articles that may be of interest to the researcher.

The system then goes deeper by analyzing the connections with the database that has abstracts, introductions, methods, and results of articles. This analysis allows the system to gain a deeper understanding of the research presented in the articles and identify the most relevant information for the researcher.

Finally, the conclusive result of this analysis is pasted to the user database. This database contains all of the relevant information and insights gained from the analysis, including relevant articles, summaries of the research presented in those articles, and any other relevant information. This user database can be used to help researchers make informed decisions about their research and gain valuable insights into their field.

To sum up we can say that this approach provides researchers with a comprehensive and efficient method for analyzing academic articles and gaining insights from the existing literature. By utilizing LLMs and advanced data analysis techniques, the system can quickly identify relevant articles and provide researchers with a clear path forward for their research.

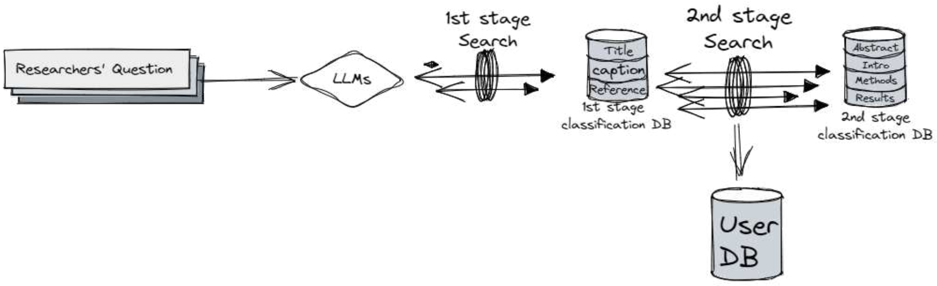

Figure 2. Analysis of Researchers' Questions among text-based databases.

4. Data-Based Search

In addition to the text analysis capabilities of the DataDiscoveryLab system, researchers' questions can also be tremendously helpful in analyzing the data. After retrieving researchers' data from their scientific software and devices, the system uses computer vision algorithms to detect and extract individual figures. The text data associated with these figures are then used to provide context, including figure captions and references, which can help researchers gain a deeper understanding of the research presented in the figures.

This analysis is particularly valuable for researchers working in fields that rely heavily on visual data, such as biology, chemistry, and physics. By extracting and analyzing the figures, researchers can gain valuable insights into the data and the conclusions presented in the articles. The conclusive result of this analysis is then added to the user database, which contains all of the relevant information and insights gained from the analysis. This information can be used to help researchers make informed decisions about their research and gain a deeper understanding of their field.

Altogether, the use of computer vision algorithms to analyze visual data provides researchers with a powerful tool for gaining insights from academic articles. By combining this analysis with NLP-based analysis of the text data and researchers' questions, researchers can gain a more complete understanding of the research presented in the articles and make informed decisions about their research. The DataDiscoveryLab system is an exciting development in the field of research analysis and has the potential to revolutionize the way researchers approach their work.

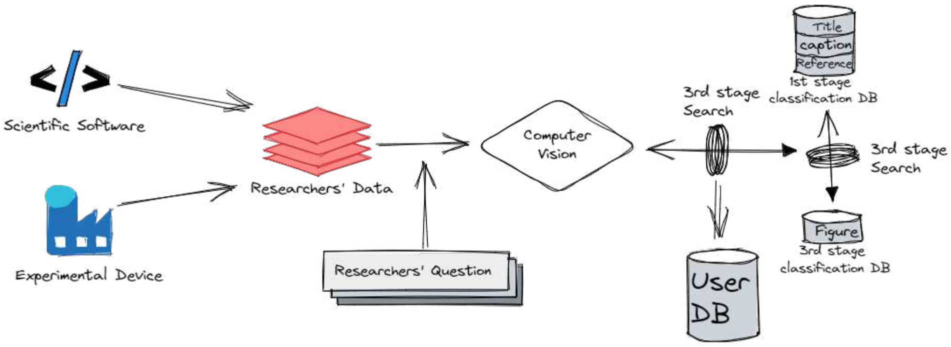

Figure 3. Analysis of Researchers' Data with Questions among text and figure-based databases.

5. Loop of Excellence

By utilizing the database that contains abstracts, introductions, methods, and results of articles, along with the user database that has been fed with the results of researchers' prompts and data, the DataDiscoveryLab system can continuously refine and optimize its analysis. The system uses LLMs to analyze the data and identify patterns and connections between the different sources of information.

This analysis is then refined according to the database that contains abstracts, introductions, methods, and results of articles, allowing the system to gain a deeper understanding of the research presented in the articles and the insights gained from the user database. This iterative process continues until the system finds the best outcome for a possible pathway to research and articles to read.

This entire process is named the Loop of Excellence, as it represents a continuous cycle of refinement and optimization that is designed to provide researchers with the best possible insights and outcomes. By leveraging advanced data analysis techniques and machine learning algorithms, the DataDiscoveryLab system can provide researchers with a powerful tool for gaining insights from academic articles and advancing their research.

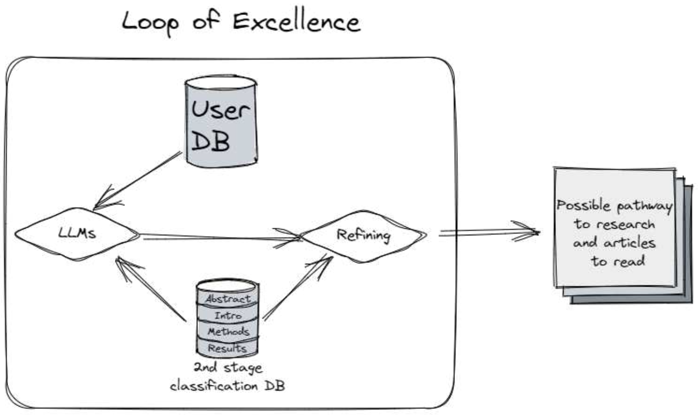
Loop of Excellence
User DB
LLMs
Refining
Abstract
Intro
Methods
Results
2nd stage classification DB
Possible pathway to research and articles to read

Figure 4. Usage of Users' Databases for nucleating recommendations.

6. Conclusion

As a result, the DataDiscoveryLab software system is a powerful tool that provides researchers with a comprehensive approach to analyzing academic articles. By utilizing computer vision algorithms and NLP techniques, the system can quickly parse articles into text and figures, creating three separate databases that can be used to find similarities between users' research questions and existing literature. Additionally, the system can integrate users' data from scientific software and devices, creating a loop that can lead to the best answer and pathways for research and articles to read.

The creation of these databases, along with the use of advanced data analysis techniques, provides researchers with a clear path forward for their research and helps them make informed decisions about their work. With the ability to parse large volumes of data and find relevant information quickly, this system can significantly enhance the research process, ultimately leading to more impactful research outcomes.

In general, DataDiscoveryLab is a valuable tool that has the potential to revolutionize the way researchers approach analyzing academic articles. With the ability to quickly and efficiently parse large volumes of data, this system can help researchers gain valuable insights from existing literature, saving them time and resources while ultimately leading to more impactful research outcomes.

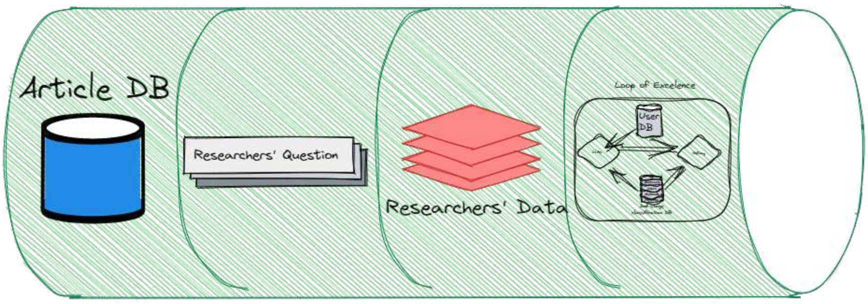

Figure 5. Entire Pipeline.

Appendix A Methods

Appendix A.1 Creating Text-Based Article Database

To create a comprehensive article database, the DataDiscoveryLab software system utilizes IBM's DeepSearch to retrieve text-based information such as references, abstracts, introductions, methods, and results. This information is then parsed and stored in the appropriate sub-databases. Additionally, web scraping techniques are used to retrieve titles and affiliations.

DeepSearch has been previously used in various publications such as [1–4]. These publications demonstrate the effectiveness of DeepSearch in extracting and analyzing text-based data.

The article database and sub-databases allow researchers to quickly identify relevant articles and gain a deeper understanding of the research presented in those articles. By utilizing DeepSearch and other techniques for data retrieval, the DataDiscoveryLab software system provides a comprehensive approach to analyzing academic articles.

After analyzing articles to parse them into text-based versions, the system uses natural language processing (NLP) methods to clean and then embed them with tools like OpenAI text-embedding-ada-002 [5]. This ensures that the text-based information is accurate and easily searchable.

Appendix A.2 Creating Figure-Based Article Database

To create a comprehensive article database, the DataDiscoveryLab software system utilizes a range of tools and techniques for extracting and analyzing scientific figures. One such tool is Deepfigures-open, which is based on the article [6]. Deepfigures-open has been widely used in other research projects, including [7–12].

The DataDiscoveryLab software system also leverages other resources to enrich the figure-based article database. For instance, the software uses EXSCLAIM, based on the article [13], to extract labeled images from articles. Additionally, the system employs EXACT, based on the article [14], to facilitate collaborative annotation of images. Other tools, such as ArtPop, based on the article [15], and TEMExtraction, based on the article [16], are also can be used to enhance the figure-based database. The MedICaT dataset, based on the article [16] which includes medical images, captions, and textual references, can be used in conjunction with other resources to enhance medical image analysis and natural language processing tasks.

The DataDiscoveryLab software system integrates these various tools and techniques into a single framework called Look, Read, and Enrich, which is based on the article [17]. This framework allows researchers to quickly search, browse, and explore the figure-based database to gain insights into the research presented in the articles. FigureSeer, based on the article [18], is also used to automatically parse result figures in research papers and add them to the database.

Overall, the creation of the figure-based article database, along with the text-based and sub-databases, provides researchers with a powerful tool for exploring and analyzing academic articles. The use of advanced NLP and computer vision algorithms, coupled with the integration of multiple tools and resources, allows researchers to quickly and efficiently extract insights from large volumes of data.

References

1. Lin, C.; Wang, P.H.; Hsiao, Y.; Chan, Y.T.; Engler, A.C.; Pitera, J.W.; Sanders, D.P.; Cheng, J.; Tseng, Y.J. Essential step toward mining big polymer data: Polyname2structure, mapping polymer names to structures. *ACS Applied Polymer Materials* **2020**, *2*, 3107–3113.
2. Manica, M.; Auer, C.; Weber, V.; Zipoli, F.; Dolfi, M.; Staar, P.; Laino, T.; Bekas, C.; Fujita, A.; Toda, H.; Hirose, S.; Orii, Y. An Information Extraction and Knowledge Graph Platform for Accelerating Biochemical Discoveries. *arXiv e-prints* **2019**, p. arXiv:1907.08400.

3. Dognin, P.L.; Melnyk, I.; Padhi, I.; Nogueira dos Santos, C.; Das, P. DualTKB: A Dual Learning Bridge between Text and Knowledge Base. *arXiv e-prints* **2020**, p. arXiv:2010.14660.
4. Staar, P.W.; Dolfi, M.; Auer, C.; Bekas, C. Corpus Conversion Service. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018.
5. Greene, R.; Sanders, T.; Weng, L.; Neelakantan, A. New and improved embedding model. Retrieved from https://openai.com/blog/new-and-improved-embedding-model, 2022.
6. Siegel, N.; Lourie, N.; Power, R.; Ammar, W. Extracting Scientific Figures with Distantly Supervised Neural Networks. *arXiv e-prints* **2018**, p. arXiv:1804.02445.
7. Ammar, W.; Groeneveld, D.; Bhagavatula, C.; Beltagy, I.; Crawford, M.; Downey, D.; Dunkelberger, J.; Elgohary, A.; Feldman, S.; Ha, V.; Kinney, R.; Kohlmeier, S.; Lo, K.; Murray, T.; Ooi, H.H.; Peters, M.; Power, J.; Skjonsberg, S.; Wang, L.L.; Wilhelm, C.; Yuan, Z.; van Zuylen, M.; Etzioni, O. Construction of the Literature Graph in Semantic Scholar. *arXiv e-prints* **2018**, [arXiv:cs.CL/1805.02262]. doi:10.48550/arXiv.1805.02262.
8. Bhatt, J.; Hashmi, K.A.A.; Afzal, M.Z.; Stricker, D. A Survey of Graphical Page Object Detection with Deep Neural Networks. *Applied Sciences* **2021**, *11*, 5344. doi:10.3390/app11125344.
9. Wang, Q.; Li, M.; Wang, X.; Parulian, N.; Han, G.; Ma, J.; Tu, J.; Lin, Y.; Zhang, H.; Liu, W.; Chauhan, A.; Guan, Y.; Li, B.; Li, R.; Song, X.; Fung, Y.R.; Ji, H.; Han, J.; Chang, S.F.; Pustejovsky, J.; Rah, J.; Liem, D.; Elsayed, A.; Palmer, M.; Voss, C.; Schneider, C.; Onyshkevych, B. COVID-19 Literature Knowledge Graph Construction and Drug Repurposing Report Generation. *arXiv e-prints* **2020**, [2007.00576]. doi:10.48550/arXiv.2007.00576.
10. Li, M.; Cui, L.; Huang, S.; Wei, F.; Zhou, M.; Li, Z. TableBank: A Benchmark Dataset for Table Detection and Recognition. *arXiv e-prints* **2019**, [1903.01949].
11. Zhong, X.; ShafieiBavani, E.; Yepes, A.J. Image-based table recognition: data, model, and evaluation. *arXiv e-prints* **2019**, [1911.10683]. doi:10.48550/arXiv.1911.10683.
12. Li, M.; Xu, Y.; Cui, L.; Huang, S.; Wei, F.; Li, Z.; Zhou, M. DocBank: A Benchmark Dataset for Document Layout Analysis. *arXiv e-prints* **2020**, [2006.01038]. doi:10.48550/arXiv.2006.01038.
13. Schwenker, E.; Jiang, W.; Spreadbury, T.; Ferrier, N.; Cossairt, O.; Chan, M.K.Y. EXSCLAIM!–An automated pipeline for the construction of labeled materials imaging datasets from literature. *arXiv e-prints* **2021**, *arXiv:2103.10631*. doi:10.48550/arXiv.2103.10631.
14. Marzahl, C.; Aubreville, M.; Bertram, C.A.; Maier, J.; Bergler, C.; Kröger, C.; Voigt, J.; Breininger, K.; Klopfleisch, R.; Maier, A. EXACT: a collaboration toolset for algorithm-aided annotation of images with annotation version control. *Scientific Reports* **2021**, *11*, 4343. doi:10.1038/s41598-021-83827-4.
15. Greco, J.P.; Danieli, S. ArtPop: A Stellar Population and Image Simulation Python Package. *The Astrophysical Journal* **2022**, *941*, 26. doi:10.3847/1538-4357/ac75b7.
16. Subramanian, S.; Wang, L.L.; Mehta, S.; Bogin, B.; van Zuylen, M.; Parasa, S.; Singh, S.; Gardner, M.; Hajishirzi, H. MedICaT: A Dataset of Medical Images, Captions, and Textual References. *arXiv e-prints* **2020**, *arXiv:2010.06000*, [2010.06000]. doi:10.48550/arXiv.2010.06000.
17. Gomez-Perez, J.M.; Ortega, R. Look, Read and Enrich. Learning from Scientific Figures and their Captions. *arXiv e-prints* **2019**, *arXiv:1909.09070*.
18. Siegel, N.; Kim, K.; Wallach, H.; Zoeller, A.; Peng, L.; Horvitz, E.; Seltzer, M. FigureSeer: Parsing Result-Figures in Research Papers. Computer Vision – ECCV 2016; Leibe, B.; Matas, J.; Sebe, N.; Welling, M., Eds. Springer International Publishing, 2016, pp. 664–680. doi:10.1007/978-3-319-46478-7_41.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.